STAT 306 Final Report - Predicting Sleep Efficiency

Group members: Heidi Lantz (96309695), Evan Zhang (67620880), Sophia Yang (33176769)

Introduction

We used the sleep efficiency dataset: https://www.kaggle.com/datasets/equilibriumm/sleep-efficiency in Kaggle, which contains 100 observations.

The data was collected in 2021 by a research team in the UK, and was collected from the University of Oxfordshire. It was collected from a local community over a period of several months using a combination of self-reported surveys, actigraphy, and polysomnography (a sleep monitoring technique).

Our research question and motivation behind the analysis of the data:

- Which variables are most important in predicting sleep efficiency? And how do these variables relate to sleep efficiency?
- We use a forward selection process to determine which variables are most relevant for predicting sleep efficiency and build regression models based only on the variables selected. After comparing different models, we select the best model and use this model to predict the test dataset and combine it with the actual values to see how accurate our model is.
- From the model, we can develop a general idea about what factors relate to sleep efficiency and how they are correlated. Therefore, it may suggest some methods to improve sleep patterns by controlling certain factors.

We included 12 variables:
The response variable:

- Sleep efficiency: a numerical variable ranging from 0 to 1 that indicates the proportion of time in bed spent asleep

The explanatory variables are:

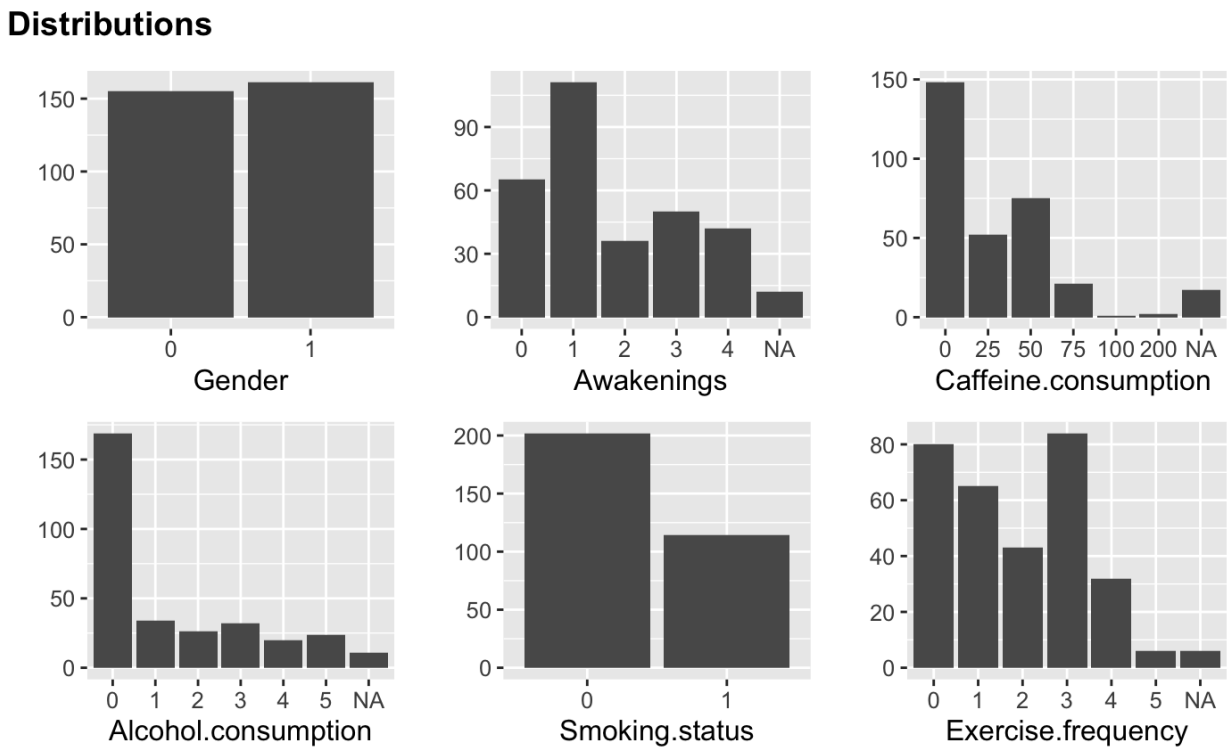
1. Age: numerical variable ranges from 9 to 69
2. Gender: categorical / dummy variable with 50% of the data is male and the rest half female
3. Sleep duration: numerical variable indicating the total amount of time the test subject slept (in hours).
4. Awakenings: categorical variable indicating the number of times the test subject wakes up during the night
5. REM sleep percentage: numerical variable indicating the percentage of total sleep time spent in REM (rapid eye movement) sleep
    - REM sleep is the stage where people have intense brain activities (dreams) and restores the areas of the brain that help with memory and learning. 20% of the total sleep time in the REM stage is considered good
6. Deep sleep percentage: numerical variable indicating the percentage of total sleep time spent in Deep (non - rapid eye movement) sleep
    - Deep sleep is important for the body to replenish energy stores and repair muscles, bones, and tissue. 15% to 25% of the total sleep time in the deep sleep stage is considered normal

7. Light sleep percentage: numerical variable indicating the percentage of total sleep time spent in Light sleep
    - Light sleep is the transitional stage between waking and sleeping. Typically take up about 50% to 60% or more of the total sleep time
8. Caffeine consumption: categorical variable indicating the amount of caffeine consumed in the 24 hours prior to bedtime (in mg)
9. Alcohol consumption: categorical variable indicating the amount of alcohol consumed in the 24 hours prior to bedtime (in oz)
10. Smoking status: categorical / dummy variable that states whether or not the person smokes
11. Exercise frequency: categorical variable indicating the number of times the person exercises each week

Although some explanatory variables seem to be numerical (e.g., Alcohol consumption, Caffeine consumption), the number of values that these variables take is limited; therefore, we can view and convert these variables into categorical with a few levels.

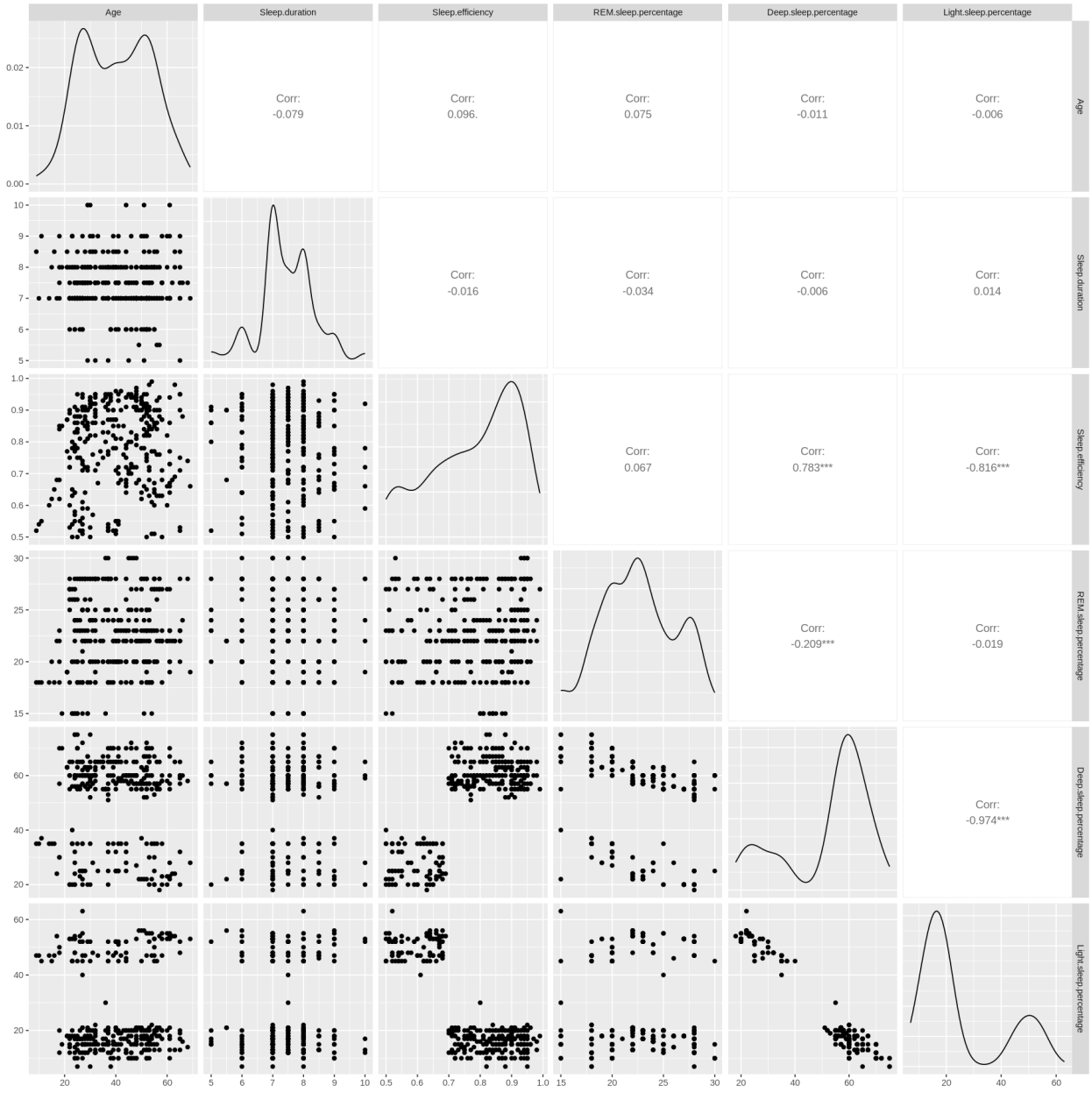
Exploratory Data Analysis

We splitted the data into test and training datasets with 70% of the data being training, and we performed the exploratory data analysis only on the training dataset.



Here we explore the distributions of all categorical explanatory variables. We can notice that we have roughly the same number of people for each gender, and there are around double the amount of

nonsmokers compared to that of smokers. Awakenings and exercise frequency are more evenly distributed, while caffeine consumption and alcohol consumption seem skewed to the right.



Looking at this ‘ggpairs’ graph which only includes the numerical variables, we see that only deep sleep percentage and light sleep percentage are strongly correlated with the response variable (sleep efficiency). Their correlations are 0.783 and -0.816 respectively.

Most of the explanatory variables have no significant correlation with each other, excluding deep sleep percentage and light sleep percentage. These two variables have a very strong correlation that is -0.974, and their corresponding scatter plot also exhibits a strong negative linear relationship. This could result in a very problematic issue called collinearity. We don’t want collinearity to affect our model and we know that taking out one of these variables shouldn’t take away too much information from the model.

Therefore, we will only include deep sleep percentage as one of the explanatory variables, and remove light sleep percentage from the data set.

Model Building

We first built a full model regression that we called Full Regression, which includes all the explanatory variables. Then we used the forward selection process and found that the best model should include 7 variables, which are Age, REM sleep percentage, Deep sleep percentage, Awakenings, Alcohol consumption, Smoking status, and Exercise frequency. We put these 7 variables into a model called Selected Regression.

We also tried to add some interaction terms to explore some potential relationships: we added interaction terms of age and other variables and further built 4 more different models. The reason we choose Age is because there are a lot of factors that may be affected by age. For example, the damage of smoking may vary for people of different ages. We want to explore how the interaction with Age would affect the model.

In summary, we built 6 model in total and obtained the following result:

Model <chr>	R_MSE <dbl>	AIC <dbl>	BIC <dbl>	adjusted_R_sq <dbl>
Full Regression	0.05625743	-770.6182	-673.3610	0.8301023
Selected Regression	0.05564806	-802.1114	-728.9218	0.8174720
Selected Regression with Interaction Age*Smoking.status	0.05590115	-803.3913	-726.5422	0.8188702
Selected Regression with Interaction Age*REM.sleep.percentage	0.05576015	-800.5412	-723.6921	0.8170625
Selected Regression with Interaction Age*Deep.sleep.percentage	0.05362911	-804.8010	-727.9519	0.8197577
Selected Regression with Interaction Age*Awakenings	0.05467584	-795.3164	-707.4888	0.8154827

6 rows

From here, we know that the model with the lowest RMSE value, lowest AIC value, lowest BIC value, and highest adjusted R squared value would give the best out-of-sample prediction performance.

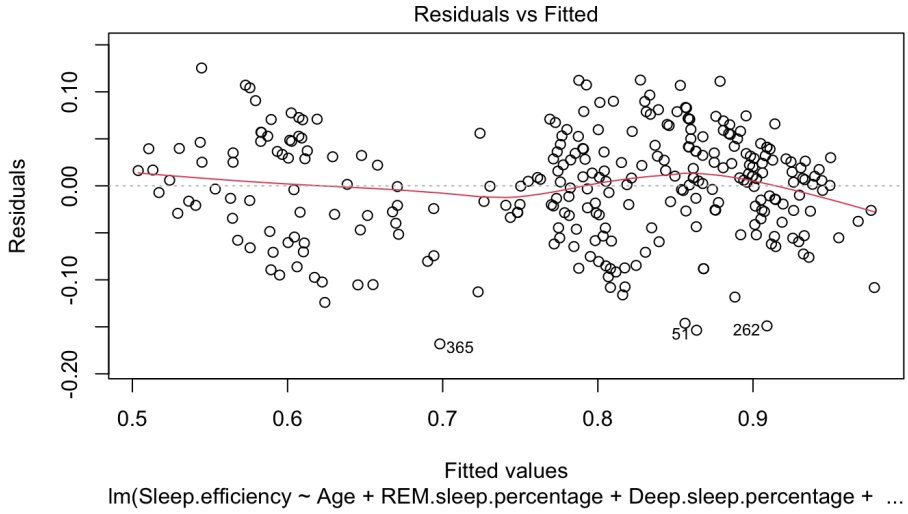
In this example, we can see that the Selected Regression model where we only selected the variables using forward regression produces a better model. The six models we build all have roughly the same value of RMSE and adjusted R square. The full regression model has slightly higher RMSE, but it also has slightly higher adjusted R square value. Therefore, we will use AIC and BIC values to compare models. Here all of the selected regression models (with or without interaction terms) have way smaller AIC and BIC values than the full regression model, but the values between each selected model do not differ much (AIC = -800, BIC = -720 approximately). Selected Regression with Interaction Age*Deep.sleep.percentage has the lowest AIC, and Selected Regression without interaction term has the lowest BIC value. Since models with less number of parameters are preferred, the Selected Regression model without interaction term is the best model. Therefore, we see that our forward selection allowed us to find the best model for us.

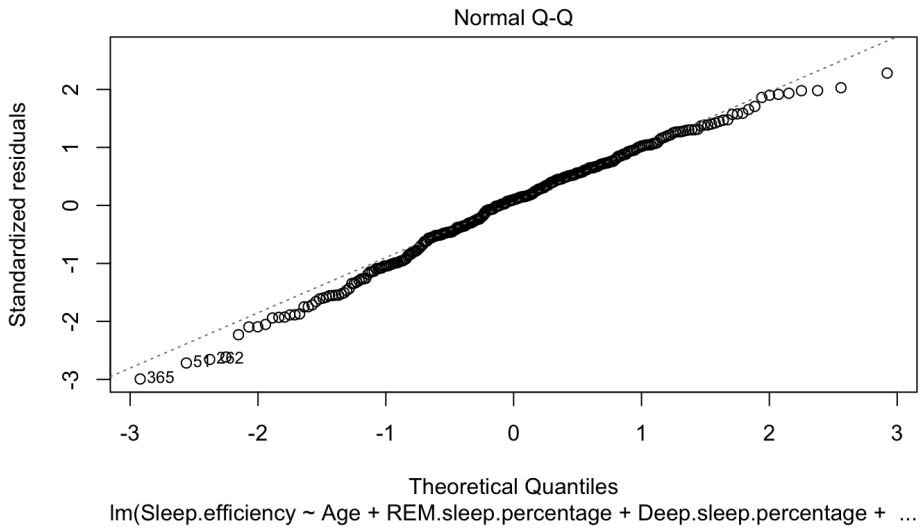
Best model (Selected Regression):

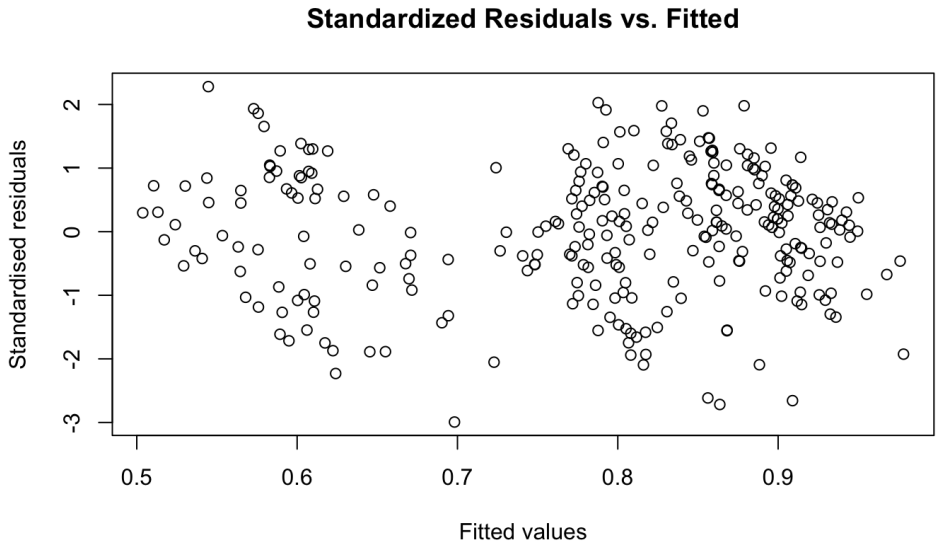
```
Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)             0.4280863  0.0354305  12.082  < 2e-16 ***
Age                     0.0009426  0.0002733   3.448 0.000655 ***
REM.sleep.percentage    0.0062495  0.0010985   5.689 3.33e-08 ***
Deep.sleep.percentage   0.0051119  0.0002837  18.020  < 2e-16 ***
Awakenings1            -0.0518292  0.0097082  -5.339 2.00e-07 ***
Awakenings2            -0.1218411  0.0135237  -9.009  < 2e-16 ***
Awakenings3            -0.1285825  0.0123698 -10.395  < 2e-16 ***
Awakenings4            -0.1292638  0.0127420 -10.145  < 2e-16 ***
Alcohol.consumption1   -0.0199864  0.0120535  -1.658 0.098460 .
Alcohol.consumption2   -0.0452141  0.0138010  -3.276 0.001191 **
Alcohol.consumption3   -0.0087290  0.0122921  -0.710 0.478244
Alcohol.consumption4   -0.0360818  0.0154111  -2.341 0.019951 *
Alcohol.consumption5   -0.0456900  0.0134628  -3.394 0.000794 ***
Smoking.status1        -0.0387113  0.0077465  -4.997 1.05e-06 ***
Exercise.frequency1    -0.0072132  0.0111471  -0.647 0.518127
Exercise.frequency2     0.0311642  0.0123306   2.527 0.012067 *
Exercise.frequency3     0.0104542  0.0099608   1.050 0.294881
Exercise.frequency4     0.0314848  0.0136130   2.313 0.021488 *
Exercise.frequency5     0.0172785  0.0277832   0.622 0.534533
```

Analysis Discussion

To see how well our model fits, we plotted the model diagnostics graphs of the Selected Regression model:







Here we see that the best model we choose may not be the most suitable. It seems that the residual plot does not have a constant variance, and the data points are clustered in 2 groups, but the residuals follow a Normal distribution approximately. The standardized residual plot also shows that we may have a problem with our model since there are too many data points lying outside the range [-2, 2]. If they follow the Normal distribution exactly, there should be only 3 data points outside this range given the empirical rule.

Therefore, we can still improve our model that we made. This can possibly be done by collecting more data and including more variables, especially changing some categorical variables into numerical. By having more information to base our model off of, we can always find ways to improve our model and help predict sleep efficiency better.

Conclusion

Based on the information above, we showed that selecting variables helps improve the model, in which we used a forward selection process to find that. We also found that in a model with 7 out of 10 selected variables: Gender, Sleep duration, and Caffeine Consumption were not as helpful in comparison to the other variables when predicting Sleep efficiency- which is an interesting discovery. It helps us to answer our research question which asks which variables are most important. We were surprised to find that caffeine consumption wasn't as important to be kept in the model, as we found many studies when researching that implied the opposite. This could be explained by the fact that we only have caffeine consumption as a categorical variable with 4 levels in our model, and hence it could be difficult to fit a linear regression on this limited data we have. If we were able to collect data as a continuous variable precise to 2 decimal places about how much caffeine people consume within 24 hours instead of 4 levels, that would help us significantly more. Then we could treat caffeine consumption as a numerical variable and fit a more accurate linear regression model based on it. We believe this would help our model's predictions and potentially change what variables are included.

To connect it back to the beginning, we remember that our research question was "Which variables are most important in predicting sleep efficiency? And how do these variables relate to sleep efficiency?". We were overall surprised to find that Caffeine Consumption was not most important. It made sense that Gender and Sleep duration weren't as important and were better to keep out of the model. We were in general happy with our model's ability to predict, however we would have liked to see a more thorough data collection on things such as alcohol and caffeine consumption, in order to better understand how these factors truly affect sleep efficiency. Our motivation in the beginning was to find out this information to help create awareness for what increases sleep efficiency. This would allow us to make recommendations on which things to focus on in order to improve your sleep. Our best-performing model suggests that seven variables help determine sleep efficiency best. We can see from the model that it is best to keep a healthy lifestyle where you avoid alcohol consumption and smoking, as well as try to prevent awakenings during your sleep. Variables such as age, exercise frequency, REM sleep percentage, and Deep sleep percentage are all things that relate positively to sleep efficiency.

In the end, we learned that a healthy lifestyle paired with a long, continuous sleep appears to produce the highest sleep efficiency. We found that our model produced by forward selection had the most effective performance in out-of-sample prediction, and in general are satisfied with how our model performed.